Digital Speech Makeup: Voice Conversion Based Altered Auditory Feedback for Transforming Self-Representation

Riku Arakawa
arakawa@star.rcast.u-tokyo.ac.jp
Research Center for Advanced Science and Technology, The University of Tokyo
Tokyo, Japan

Zendai Kashino
kashino@star.rcast.u-tokyo.ac.jp
Research Center for Advanced Science and Technology, The University of Tokyo
Tokyo, Japan

Shinnosuke Takamichi
shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
Graduate School of Information Science and Technology, The University of Tokyo
Tokyo, Japan

Adrien Verhulst
adrienverhulst@star.rcast.u-tokyo.ac.jp
Research Center for Advanced Science and Technology, The University of Tokyo
Tokyo, Japan

Masahiko Inami
inami@star.rcast.u-tokyo.ac.jp
Research Center for Advanced Science and Technology, The University of Tokyo
Tokyo, Japan

ABSTRACT

Makeup (i.e., cosmetics) has long been used to transform not only one's appearance but also their self-representation. Previous studies have demonstrated that visual transformations can induce a variety of effects on self-representation. Herein, we introduce *Digital Speech Makeup* (DSM), the novel concept of using *voice conversion* (VC) based auditory feedback to transform human self-representation. We implemented a proof-of-concept system that leverages a state-of-the-art algorithm for near real-time VC and bone-conduction headphones for resolving speech disruptions caused by delayed auditory feedback. Our user study confirmed that conversing for a few dozen minutes using the system influenced participants' speech ownership and implicit bias. Furthermore, we reviewed the participants' comments about the experience of DSM and gained additional qualitative insight into possible future directions for the concept. Our work represents the first step towards utilizing VC to design various interpersonal interactions, centered on influencing the users' psychological state.

CCS CONCEPTS

• **Human-centered computing** → **Auditory feedback**; *Interaction techniques*; • **Computing methodologies** → Machine learning.

KEYWORDS

voice conversion, auditory feedback, speech transformation, self-representation

ACM Reference Format:
Riku Arakawa, Zendai Kashino, Shinnosuke Takamichi, Adrien Verhulst, and Masahiko Inami. 2021. Digital Speech Makeup: Voice Conversion Based Altered Auditory Feedback for Transforming Self-Representation. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479934

1 INTRODUCTION

Transforming one's visual appearance is a universal human activity that has persisted through generations, from ancient Egypt to the present day [15]. While often thought of as an expression of culture, changing one's visual appearance (e.g., makeup) is also known to induce powerful psychological effects, such as improving self-confidence [34]. In the late 20th century, the advent of computers and the advancement of computer graphics processing introduced new methods of transformation using digital technologies. Examples include the many camera filters that are equipped in consumer camera applications and digital avatars that embody users in digital worlds. The high level of visual expressiveness made possible by these digital technologies allows individuals to change their appearance independently of their actual body structure. This unprecedented freedom of visual expression has sparked a significant body of work on using visual transformations for modifying human self-representation, i.e., the image people have of themselves. Specifically, previous research revealed that changing one's appearance in virtual environments using avatars can affect an individual's implicit bias or negative stereotypes based on visual factors such as race [7, 21, 33, 38], gender [29, 32], and age [8, 54].

However, transformations enabled by computers are not limited to changes in the visual modality. Methods for speech transformation have also been developed through work in the field of speech

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479934

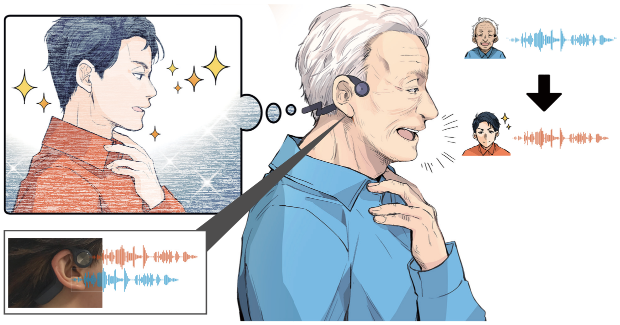

Figure 1: Digital Speech Makeup (DSM) is the concept of using real-time voice conversion (VC) as auditory feedback to transform the users' self-representation (transformation from being elderly to being young is depicted here). VC-based auditory feedback (VCAF) is transmitted through bone-conduction to avoid speech disruption due to delayed auditory feedback (DAF).

processing [45]. Since voice is an important factor in human self-representation [24], it is conceivable that speech transformations could alter the human psychological states, just as visual transformations can. Indeed, there have been several studies attempting to show the psychological effects of providing auditory feedback using speech transformations [5, 16, 27, 39, 48]. Nevertheless, exploration into the psychological effects of speech transformations is sparse when compared to studies investigating the effects of visual transformations. Specifically, existing works on this topic only utilize voice changers that modify basic speech parameters, resulting in a limited variety of transformations (e.g., affecting users' emotions). In other words, they have not taken advantage of the high-fidelity transformations afforded by recent signal processing techniques, such as *voice conversion* (VC) [45] that can completely transform the speaker's voice into another person's.

In this paper, we propose *Digital Speech Makeup* (DSM): the novel concept of using VC-based auditory feedback to transform human self-representation (see Figure 1). In detail, DSM is the concept of enabling voluntary modifications to one's self-representation using real-time VC. Namely, it is the idea of changing a user's perception of their own voice (i.e., by allowing the user to perceive their voice to be that of another person's) to change their overall perception of themselves. To achieve this, we provide VC-based auditory feedback (VCAF), which consists of the user's utterances converted to be spoken in another person's voice, to the user in near real-time while speaking. We expect successful DSM to make a positive impact on users in a variety of situations by modifying the user's self-representation through VCAF. For example, it could be used to improve vitality and liveliness in the elderly by allowing them to speak in a young voice or to become more immersed in the character of a digital avatar.

To demonstrate the efficacy of DSM, we first developed a proof-of-concept system enabling DSM. Here, we introduced the use of bone-conduction headphones to avoid speech impairment due to *delayed auditory feedback* (DAF) [53]. Then we conducted a user study and empirically showed that conversing using the system could influence users' self-representation (measured in terms of implicit bias). User comments supported our findings, qualitatively showing that they felt a change in how they perceived themselves while using DSM. Their comments further showed that there was a general trend towards being supportive of the concept of altering one's self-representation through auditory feedback. Our findings and subsequent discussion serve as a foundation for utilizing emerging VC techniques in interpersonal interactions.

2 RELATED WORK

Our work contributes to research on the relationship between computational transformations of the self and self-representation. Our literature review, thus, first covers existing studies using visual transformations, especially in VR, as these works are the most prevalent in this field. Then, we discuss studies that use speech transformations for influencing self-representation, which are sparse in contrast. Finally, to provide some background on our proposed system, we describe the state-of-the-art VC algorithms and the detrimental effects of DAF.

2.1 Visual transformations and self-representation

There is a large body of research exploring the influence of visual transformations on self-representation, especially in VR environments [13, 26, 36, 43, 55]. Many of these works investigate behavioral changes which are induced in subjects who perceive ownership of a virtual body. For example, Yee and Bailenson [55] found that people embodied in more attractive avatars in VR become more intimate with confederates than those embodied in less attractive avatars.

One well-investigated consequence of this body ownership illusion is an influenced bias and stereotyping towards the demographic that the virtual body is representative of. Yee and Bailenson [54] conducted an experiment where each participant was embodied in an avatar of either an old or young person to gain their perspective. They reported that negative stereotyping of the elderly was significantly reduced in the participants who were embodied in elderly avatars. Similar results have been reported in various visual factors which are often used to classify people, such as race [7, 21, 33, 38] and gender [29, 32]. In these studies, the degree of implicit bias was often quantitatively measured using the Implicit Association Test (IAT) [19]. Past work implies that IAT scores could be used as an indicator of whether a user's self-representation was changed due to a digital transformation[6, 9].

The aforementioned works have shown that visual transformations have a significant impact on self-representation. However, the approaches taken in the above works have yet to be applied to influence self-representation through interactions that pervade everyday life. This is because previous studies often needed to employ full-body tracking systems, such as the Microsoft Kinect or motion capturing systems, to achieve visuomotor synchrony, which plays an important role in generating the illusion of ownership over the virtual body [38, 42]. Additionally, previous studies needed to use a head-mounted display that not only provides visual stimuli to participants but also tracks their head movements with its embedded inertial measurement units.

In contrast, we propose the novel concept of achieving more habitual modification of human self-representation by making use of speech transformations instead of visual ones. Specifically, we propose to transform speech feedback such that the user sounds, to themselves, as if they are someone else. Since speech feedback is omnipresent and is significantly simpler (in terms of the hardware requirements) to generate/modify than visual feedback, it is expected that we can achieve results similar to those achieved by visual transformations in a more pervasive manner.

2.2 Speech transformations and self-representation

Unlike the many studies about the use of visual transformations as we outlined in Section 2.1, few studies have been conducted to explore the relationship between speech and self-representation [40]. There are only a few findings regarding the psychological effects of providing users auditory feedback consisting of transformed speech [5, 16, 27, 39, 48]. For example, Aucouturier et al. [5] found that the emotional state of the participants changes in congruence with real-time pitch alteration. They used three pitch patterns corresponding to happiness, sadness, and fear. The participants were asked to read sentences aloud for approximately ten minutes while hearing pitch-altered speech. To avoid impairment due to DAF, they developed specialized hardware that reduced latency to only 15 ms.

In an effort to intervene in self-representations beyond influencing users' emotion, Tajadura-Jiménez [48] et al. provided child-like voice feedback to participants who were embodied in child-like avatars in VR. The participants were asked to describe objects they saw in VR for approximately five minutes while hearing converted speech feedback. However, they found that auditory cues were not powerful enough to enhance the visual transformation. To achieve a child-like voice, they used audio editing software that changes speech parameters (e.g., by raising speech pitch by four semitones). The software introduced latency of approximately 50 ms, and it was implied that speech impairment due to DAF was not observed in their study.

As the relationship between speech transformations and self-representation has been only sparsely studied, there are some questions that have yet to be answered: *Is it possible to intervene in self-representation using audio stimuli when using high-fidelity state-of-the-art VC algorithms? Can self-representation be influenced using only speech transformations (without visual stimuli)?* In this paper, we aim to lay the groundwork for discussing the potential of speech transformations by providing preliminary answers to these questions. By doing so, we take the first step towards enabling a variety of applications by leveraging VC techniques to augment interpersonal communication.

2.3 Voice conversion techniques

Voice Conversion (VC) refers to a speech processing technique that alters one's speech into another's by changing para-/non-linguistic information while maintaining the linguistic information [44, 46, 50]. It differs from voice changers, which simply modify speech parameters such as pitch and formant, and is known for its high expressiveness and the naturalness of the converted speech [52]. One of the most popular approaches to VC is based on statistical processing, which is a data-driven approach capable of extracting a complex conversion function for transforming speech [49].

Among them, recent implementations that work in near real-time [2, 51] have been receiving attention from researchers in HCI. These algorithms sequentially convert the input speech with a real-time factor less than 1, resulting in an algorithm latency of 50 ms. This latency is a result of the recurrent processing of the algorithm and a minimum value of 50 ms is necessary to achieve high quality converted speech [2]. Use cases of these VC systems in our everyday lives have been envisioned for augmenting near-field speech communication [3]. The development of such near real-time VC systems and their suggested use in everyday interactions inspired our study to investigate the influence of VC on the users' psyche. Our work will serve as a foundation for future work seeking to utilize VC as an approach for affecting users' psychological state in social interaction, in particular, self-representation.

2.4 Delayed auditory feedback

When we try to make use of speech feedback in experiments, it is inevitable to consider the effects of delayed auditory feedback (DAF) [53]. This is a well-known human auditory characteristic: when we hear our voice with a small time delay, our speech performance may be seriously affected, resulting in unnaturally stretched vowels and stuttering. A number of speech experiments under the DAF condition have been reported [14, 30]. From their results, it is clear that auditory feedback is taken into account in some involuntary information processing in the speech motor control system. The range of delay that causes speech disturbances varies from person to person. Although a clear threshold is not known, it is empirically reported that detrimental effects from DAF occur if the delay is between 30 ms and 300 ms [10, 28].

Even using the current state-of-the-art algorithms, real-time VC incurs at least 50 ms of latency, as we mentioned in Section 2.3. Making use of real-time VC for altered auditory feedback, thus, is non-trivial as this latency lies squarely within the delay range where detrimental effects due to DAF are known to occur. In this paper, a novel approach is adopted to avoid the detrimental effects of DAF while providing VCAF. Namely, we leverage the alternate audio-transmission path used by bone-conduction headphones to minimize the effects of DAF while still providing auditory feedback. We elaborate on the system implementation in the following section.

3 DIGITAL SPEECH MAKEUP

3.1 Envisioned scenarios

We envision that DSM will be used to voluntarily alter users' self-representation in everyday life based on their demands. Figure 2 shows several examples of future DSM usage scenarios. Figure 2(a) shows an elderly person feeling more energetic by hearing themselves speak with a young voice through VCAF. Figure 2(b) shows a scene where a group of people are conversing while hearing VCAF in a voice of their choice such that they can converse in the mood of their favorite personality. In addition to being used in everyday communications like these, DSM will be appreciated by those who are looking to act as a different persona in special circumstances. For example, Figure 2(c) shows a performer being able to act as

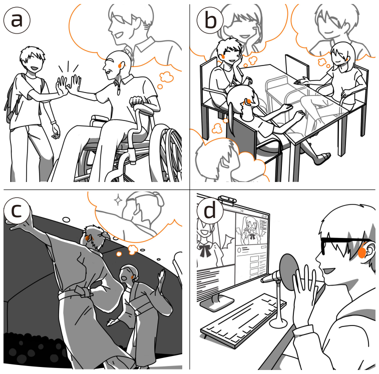

Figure 2: Future scenarios of Digital Speech Makeup (DSM). DSM can be utilized not only in everyday communications (a, b), but also in specific situations where users are looking to act as a different persona (c, d).

if they became their ideal actor through the assistance of VCAF. Figure 2(d) shows a live streamer immersing themselves in the role of their original character by hearing their speech converted to that of the character via VCAF.

3.2 Proof-of-Concept implementation

The overall system developed for our proof-of-concept is shown in Figure 3. We implemented the VC algorithm presented by Arakawa et al. [2] as it is able to generate high-quality output speech with minimal latency. Although their algorithm assumes one-to-one conversion (i.e., the model converts a given speaker to a given target speaker), we trained any-to-one conversion models using pre-recorded multi-speaker speech data. In other words, the VC system using the trained model can convert any speakers' speech to the desired target speaker's voice [47]. To achieve this, we first chose some men as source speakers from a public corpus. Then, we trained a model using the speech data of the chosen speakers as source data. We then repeated the process with several women speakers. Consequently, we obtained two separate models: one for converting from a male voice to the target voice, and for converting from a female voice to the target voice. When we had our participants try the system in the following user study, we selected the model based on their biological sex. This allowed us to conduct experiments using VCAF without recording each participant's speech data.

We used a conventional Linux PC (Intel (R) Core (TM) i7-3770K CPU @ 3.50 GHz) to run the real-time VC (Figure 3 (a)). The end-to-end latency was approximately 200 ms, which is a sum of the algorithmic latency (i.e., 50 ms) and other sources of latency (e.g., processing, I/O, communication). We also used a conventional condenser microphone that recorded users' speech at a sampling rate of 16 kHz.

In addition, to enable users to continue speaking without speech impairment due to DAF when they heard VCAF with 200 ms delay, we used a pair of off-the-shelf bone-conduction headphones connected to the PC via Bluetooth to transmit VCAF to the user (Figure 3 (b)). The idea is to transmit the converted speech to the auditory system through the path of bone conduction without blocking the ear canal. This allows us to keep the ear canal open to receive vocal auditory feedback that is coming through the air. By doing this, we expected that the users would be able to speak normally, relying on their original voice for speech motor control, while simultaneously receiving VCAF with a small delay. To verify this, we conducted a pilot test and confirmed that users can speak normally without encountering the speech impairment due to DAF. We further anticipated that VCAF that is provided to users by taking such an approach would affect their perception regarding their speech, and eventually induce psychological effects on their self-representation, which we will evaluate in Section 4.

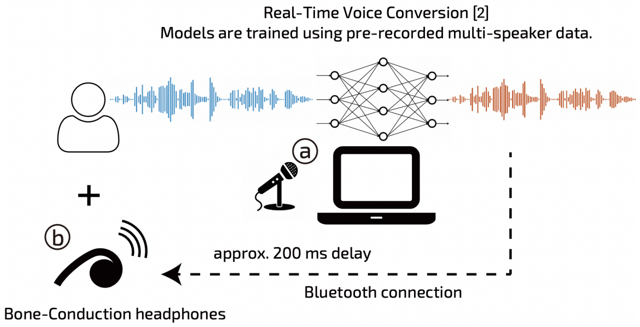
Real-Time Voice Conversion [2]
Models are trained using pre-recorded multi-speaker data.
+
a
b
approx. 200 ms delay
Bluetooth connection
Bone-Conduction headphones

Figure 3: The overview of the proof-of-concept system.

Note again that the system provides VCAF only to the speaker, not to any listeners. This means that listeners do not hear the converted speech but hear the speech before it is converted. We chose to only transmit the VCAF to the speaker as we focused on influencing the psychological states of the users (i.e., the speaker) as the first step of DSM.

4 USER STUDY

Using the proof-of-concept system described above, we investigated how human psychological states can be influenced by experiencing DSM. Specifically, we sought to identify whether the system can induce psychological effects (i.e., changes in self-representation) similar to those observed with visual transformations, as we discussed in Section 2.1. Inspired by prior studies investigating the effects of visual transformations, we chose to evaluate whether transformations in self-representation had occurred by examining how the users' implicit biases changed as a result of using our system.

4.1 Design

As we mentioned in Section 2.1, many prior studies have been conducted to examine the effect of visual transformations on the users' implicit bias toward different groups of race, age, and gender. To assess the effects of our proof-of-concept system, we decided to focus on investigating the effect of speech transformations on bias toward elderly people, a group often used to examine the efficacy of transformations [8, 54]. Our investigation was carried out by recruiting young participants and having them experience real-time VCAF with an elderly voice, using a within-participant design. Specifically, the user study included 17 participants (P1 ~

P17) without compensation aged between 21 to 38, of whom 6 were women. All participants were fluent speakers of Japanese.

As we discussed in Section 3, the proposed system is envisioned to be used in our daily lives, such as in everyday conversation. Therefore, to replicate such situations, the task we gave the participants was to have a conversation for a few dozen minutes while using the proof-of-concept system. We then evaluated the effects of this intervention both quantitatively and qualitatively.

4.2 Material

The VC models we used in this study were trained beforehand to convert speech of young speakers to that of the elderly. To do this, we used the JNAS database [23] because it contains data of Japanese from various age groups. From the database, we chose a man speaker "M049" and a woman speaker "F105" as target elderly (over 60 years old) speakers. We chose two men and two women as source speakers from the JNAS database and our in-house databases. The training process was the same as described in Section 3.2. In the experiment, when the participant was a man, we used the model converting a young man's voice to that of an elderly man. Similarly, when the participant was a woman, we used the model converting a young woman's voice to that of an elderly woman. Note that, since the VC algorithm is adopted from [2] and its transformation quality was already confirmed, we did not formally evaluate our model's quality in this paper. Instead, we asked five people (none of them were from the participants) to listen to some sample converted phrases through the bone-conduction headphones. As a result, they all agreed that the speech sounded like the elderly's voice, indicating that the transformation was suitable for our proof-of-concept system.

4.3 Measures

4.3.1 Implicit Association Test. Following works on visual transformation (Section 2.1), we focused on evaluating the implicit effect of DSM. Specifically, we followed [6, 9, 32] and used the Implicit Association Test (IAT) to measure participants' implicit bias towards age [19, 20]. In this test, participants are required to sort pictures quickly. Each picture comes from one of two categories, such as male and female. At the same time, participants also sort words into one of two categories, such as positive ones and negative ones. The IAT score can reflect the strength of the participant's implicit association between each pair of the picture category and the word category. The score can be calculated based on the latency and accuracy of their sorting performance based on the procedure described in [20].

In our study, as we aimed to measure the bias toward the elderly, the task was to classify some pictures of young and elderly faces as well as positive and negative Japanese words. A higher positive IAT score implies higher implicit bias towards associating the elderly with negative words and the young with positive words. Past studies investigating the effects of visual transformations on self-representation reported a change in the score after participants were transformed to have different visual appearance [8, 21]. Since we anticipated that DSM could induce a similar effect on users' bias, we expected that the participants' IAT score would similarly be changed after using the proposed system.

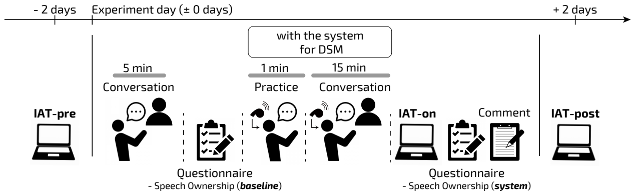

Figure 4: Illustration of procedure of the user study.

4.3.2 Speech ownership. In addition to the IAT, we evaluated speech ownership to measure how participants perceived the converted speech feedback. Sense of ownership is a measurement frequently used in experiments involving visual transformations to assess the level of embodiment [11, 31]. Sense of ownership is typically measured by collecting subjective ratings from participants using a Likert scale, analyzing the results, and comparing them over several experimental conditions [18]. The following two questions are often used to measure speech ownership [40, 56]: "I felt as if the voice I heard was my own voice" (Q1) and "I felt as if the voice I heard was a modified version of my own voice" (Q2)

We adopted the same questionnaire, that is, the above two questions (Q1 and Q2) using a 7-point Likert scale (with 1 indicating "strongly disagree" and 7 indicating "strongly agree"). We introduced these questions to examine whether participants perceived their voice being modified when using the system, compared to the usual utterance without the system. Here, we expected that there would be significant differences between the scores to these questions between the conditions without the system (*baseline*) and with the system (*system*).

4.4 Procedure

The procedure of the user study is shown in Figure 4. All participants had their first IAT two days before the experiment day (IAT-pre). At this point, they also read and agreed to the research policy. On the day of the experiment, the participants first had a simple conversation with the experimenter without using the system for approximately five minutes. Then, they were asked to fill out the questionnaires for speech ownership to evaluate the normal conversation experience without the system (*baseline*). Next, they wore the bone-conduction headphones and adjusted the volume to be comfortable for them. Then, they started to experience the elderly VCAF produced by the model we had trained. Here, we had an initial one-minute practice session before proceeding to the main experiment. In the practice session, the participants got accustomed to speaking while being aware of the elderly VCAF. After this session, each participant conversed with the experimenter for 15 minutes. During the conversation, the participant was asked to answer random questions such as "Tell me about your favorite book". Here, the experimenter focused on listening more than speaking such that the participant could speak more, and thus, have greater exposure to VCAF. Immediately after this conversation session, we removed the bone-conduction headphones and asked the participants to perform their second IAT (IAT-on). After finishing the test, they filled out the questionnaires for measuring speech ownership

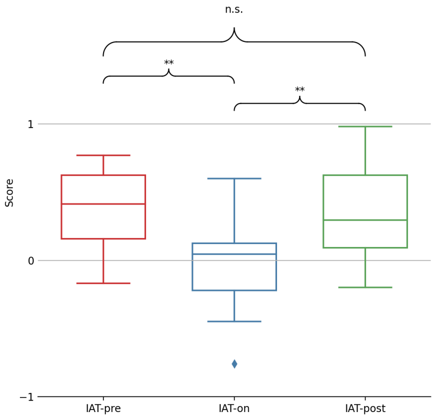

Figure 5: Participants' scores for IAT tests. n.s. – not significant, $^{**}p < 0.01$. A higher positive IAT score implies higher implicit (negative) bias towards elderly. The results indicate that the participants' negative bias towards the elderly was reduced immediately after they used the system (IAT-on).

(*system*). Then, the participants were asked to provide some comments on the experience, including feelings they had and anything they noticed during the session. Finally, each participant took their third IAT two days after the session (IAT-post).

4.5 Result

4.5.1 IAT. A one-way repeated measures ANOVA was conducted on the participants' IAT scores to determine the significance of the results. The analysis was significant, $F(2, 48) = 11.3232$, $p = 0.0002$. We then conducted a post-hoc test using the Tukey-Kramer test and noted that there were significant differences between the IAT-pre and IAT-on results ($p = 0.002$), and the IAT-on and IAT-post results ($p = 0.005$). In contrast, there was no significant difference between the IAT-pre and IAT-post results ($p = 0.9$). Figure 5 shows participants' scores for each IAT test.

The results showing a significant difference between the IAT-on and IAT-pre scores suggest that the experience of using DSM can influence implicit bias (reduction toward elderly people in our case). Furthermore, the significant difference between the IAT-on and IAT-post scores and the insignificant difference between the IAT-pre and IAT-post scores suggest that the effect of makeup did not last long after the participants experienced it. In summary, the results indicate that the proposed system for DSM can successfully influence the user's self-representation while being used. We discuss the possibility of the long-term effect of DSM later in Section 6.1.

4.5.2 Speech ownership. Table 1 shows a summary of the subjective ratings given by participants in response to the two questions on speech ownership, Q1 and Q2, respectively. In both questions, a paired t-test showed that there were significant differences between the scores for *baseline* and *system* conditions. In detail, the degree that they felt the heard voice was their own was significantly reduced with the system, while the degree that they felt the heard voice was a modified version of their speech was increased. This result is aligned with our expectation that the participants, at least, perceived that their voice was altered while they were holding a conversation with VCAF.

Table 1: Comparison of speech ownership between the *baseline* and *system* conditions. The questionnaire for speech ownership consists of two questions: Q1 corresponds to the degree of perceiving the speech as their own speech while Q2 corresponds to the degree of perceiving the speech as a modification of their own speech.

Measure	*baseline*	*system*	p-value
Speech ownership (Q1)	6.94 (±0.24)	4.13 (±1.17)	< 0.05
Speech ownership (Q2)	1.13 (±0.33)	4.56 (±1.22)	< 0.05

4.6 User comments

The above results indicate that the proposed system was able to influence the user's speech ownership and self-representation in terms of their implicit bias. In this section, we review comments obtained from participants at the end of the experiment to further explore our results qualitatively.

Overall, the participants favored the experience and found it interesting. In detail, twelve of the participants mentioned finding the experience of hearing someone else's speech feedback while speaking to be fun. Moreover, ten of them mentioned that they felt the illusion of becoming someone else during the experience.

> I felt as if I were an old man during the conversation, which was an interesting and new experience. (P1)

Moreover, five of the participants reported a gradual and unconscious transition regarding the strength of the illusion. In other words, they did not feel the change in their self-representation at the beginning of the conversation, but they found themselves feeling more elderly by the end.

> At first, I was a little confused about the new experience and I felt I was paying more attention to my original speech. However, after several minutes of speaking, I became accustomed to the experience and, at that moment, I felt myself slightly change to be someone else. (P2)

> Gradually, I became unaware of which of the two voices I was hearing was mine. (P5)

Interestingly, one of the participants reported that they noticed an unconscious behavioral change in the latter part of the conversation.

> I found myself talking while bending at the waist like an old man. Considering this, I was really into the converted speech. (P9)

Eight participants mentioned their desire to use other VC models in their daily lives, which supports the concept of DSM.

> I want to use this system in daily life to change my voice freely. I'd like to change my voice to [name of a Japanese actor] because, then, I may be able to feel like him. (P4)

> I'm wondering what kind of feeling I would get if I use the system to change my voice to a woman's. It's an exciting thought. I might be able to chat with girls more intimately if I feel that way. (P7, man)

In addition, two participants who reported prior experience using speech-altering technologies gave favorable comments regarding their experience with the system for DSM. For example, one participant commented as follows.

> I regularly stream live as a VTuber[1]. I change my voice by modifying pitch with an off-the-shelf voice changer. However, during live-streaming, I cannot hear the converted speech feedback due to speech impairment caused by DAF. This means that only the listeners can hear the converted speech. In this regard, I found the bone-conduction system very useful in that I can check the converted speech while speaking, and what's more, I can be more immersed in my character, which is what I've been wanting for! (P10)

At the same time, however, there were several comments that demanded further improvements to the system, especially regarding the latency.

> It was a bit strange for me to hear speech feedback with delay. (P3)

These comments illustrate the qualitative aspects of the experience using DSM. We will discuss the findings in the next section.

5 DISCUSSION

5.1 Handiness of DSM

As we described in Section 3, our concept of DSM envisions pervasive use of speech transformations to voluntarily modify self-representation. We expected that the handiness of DSM would be appreciated since speech feedback can be omnipresent and speech transformation can be carried out without requiring many external devices. The concept is supported by the participants' comments regarding the desire to use the system habitually in their daily lives, as we mentioned in Section 4.6.

Furthermore, our result of reducing implicit bias in a handy manner can be utilized in more specific situations, too, such as training and education. For example, the demand for utilizing VR technologies as a training method to reduce bias at the workplace has been increasing [17, 37]. However, simultaneously, it is pointed out that one disadvantage adherent to such training approaches is the time and cost of using the system [37]. This is because they assume visual transformations to induce perspective-taking and, thus, require use of external devices, as we discussed in Section 2.1. If the illusion of being somebody else can be achieved through speech transformation, our approach can mitigate the issue by offering a handy approach. For example, given the results of the study presented herein, our system could be harnessed if a company aims to introduce a training program for reducing employees' bias toward the elderly. Although further investigation is desirable to design such a training program, we believe that our work will open up a variety of applications leveraging the handiness of DSM.

5.2 Mindlessness of DSM

The obtained comments implied that the transition of the strength of the illusion (i.e., feeling the self as elderly) occurred in a gradual and unconscious manner, as we mentioned in Section 4.6. While making transformations unconscious was not the primary intention of the system, it is a welcome feature as it implies the method of transformation poses no additional cognitive load. Furthermore, it provides another way to look at our work. Namely, it suggests that DSM can be examined from the perspective of mindless computing, an approach for designing behavior-changing technologies that leverage human biases or unconscious behaviors [1].

When examined from a mindless computing perspective, DSM could be considered to be similar to the Mindless Attractor, a method for unconsciously drawing human attention by perturbing parameters of speech [4]. Advantages from works based on mindless computing like the Mindless Attractor, thus, likely apply to DSM as well. For example, it is known that mindless interventions can be effective regardless of the users' motivation level or attitude. This would be particularly beneficial in some of the scenarios we mentioned in Section 5.1, such as training or education, where it is difficult to assume that users are always motivated to change.

5.3 Desire to use a variety of VC models

In our study, we trained and used a VC model that converts young people's voice to that of the elderly. By doing so, we showed that the proposed proof-of-concept system can influence young people's bias toward the elderly. This was our first step towards freely influencing human self-representation and demonstrates the potential of DSM. However, the concept of DSM itself is broader than this result in that it envisions a conversion of any voice to any other, and this is why we utilized VC algorithms, as we described in Section 3. During our study, we found that a desire for this kind of system was not only limited to the authors. For example, one of the participants (P4) commented that they wanted to convert their voice to that of their favorite actors. Therefore, other variations of VC as makeup are desirable to be investigated, including the example of a training program at the workplace we mentioned in Section 5.1. We believe our work will serve to spur on such investigations by demonstrating that psychological changes can be induced by utilizing VCAF in combination with bone-conduction headphones to avoid speech impairment due to DAF.

6 FUTURE WORK

Some points remain to be explored in future work to further contribute to research on transforming the self and its associated effects.

6.1 Long-term and side effect of DSM

The IAT scores we obtained suggest that the effect on the user's implicit bias does not last for a few days after the experience, as we discussed in Section 4.5.1. A similar persistence of the illusion of transformations has been observed in the studies utilizing visual transformation. For example, Banakou et al. [7] empirically confirmed that the reduced racial bias caused by visual transformations would last one week by measuring IAT scores, and Herrena et al. [22] showed that there were visible effects of empathy toward the homeless after a VR "perspective-taking task" for at least eight

[1] A **v**irtual You**Tuber**, an online entertainer who is represented by a digital avatar.

weeks. In these works, the results implied that the number of exposures to the transformations correlated with the strength of the caused illusion. Here, as discussed in Section 5.1, the handiness of DSM will motivate users to frequently utilize the makeup in everyday life. Thus, it is expected that the frequency of exposures using the proposed system will be much higher than past works have investigated. As such, future work investigating the effects of DSM will include queries into the long-term effects of unprecedentedly persistent transformations (e.g., using DSM for a week).

In addition to potential long-term psychological effects, we expect that side-effects of persistent transformation may occur in user speech patterns. There are studies regarding the effects of hearing altered speech feedback on speech itself [30]. In detail, it is well-known that humans will adjust vowel sound production in subsequent utterances when they receive altered speech as feedback [12, 25]. It's also known that such a change occurs after participants are exposed to speech feedback for weeks [35]. Such potential effects of DSM on users' speech production must also be investigated quantitatively in future experiments.

6.2 Visual v.s. speech transformation

In addition to long-term effects, our work will also spur investigations into the effect of speech transformations in combination with visual transformations. With respect to this, as mentioned in Section 2.2, Tajadura-Jiménez et al. [48] reported that speech transformation that outputs a child-like voice using a voice changer did not increase the illusion of transformation when the participants were embodied in child-like avatars. Also, they showed that the strength of the embodiment illusion was diminished if the child-like voice feedback was incongruent with the appearance of the virtual body. Considering their result, further experiments with more participants are required to investigate the relationship of multimodal transformation of the self. We believe that our proposed approach of using VC instead of voice changers can contribute to such studies by offering a larger variety of speech transformations.

7 LIMITATION

While also its fundamental advantage, a fundamental limitation of DSM is its exclusive use of VC. Specifically, the variation of the makeup is technically constrained to the expressiveness of human speech. In other words, the proposed system is unlikely to be leveraged to transform self-representation in those who do not have speech characteristics unique enough for users to distinguish the group the speech represents. We should note this in further investigating application scenarios, for example, training programs toward reducing bias at the workplace as we mentioned in Section 5.1. This also motivates us to conduct explorations into the relationship of multimodal transformation and self-representation so as to expand the scenarios, as we discussed in Section 6.2.

Next, we describe some limitations existing in the current proof-of-concept system. First, we need speech data of a person that we want our speech to be converted to. Specifically, our current implementation of VC, which followed [2], relies on parallel speech data, which means both the source and target speakers need to utter the same sentences. It usually takes hours for a person to record such utterance data in a quiet place. This may hinder the future prevalence of DSM. Notably, there has been investigation into techniques that make achieving VC easier. In the field of signal processing, for example, there are algorithms that use non-parallel data to achieve VC [41]. Future work can incorporate these techniques to mitigate the burden of recording data.

Secondly, some participants commented on the unnaturalness of latency caused by the system during the user study (Section 4.6). As we mentioned in Section 3.2, the current system has 200 ms latency, which can be reduced if we implement the system with lower-level programming (e.g., by using an FPGA). It is noteworthy, however, that existing VC algorithms have an inherent 50 ms delay due to the recurrent processing used in the algorithm, as we discussed in Section 2.3 and Section 3.2. Thus, we believe that reducing the system latency can be effective for offering better user experiences, but that approaches for avoiding speech impairment due to DAF are still necessary. Utilizing bone-conduction headphones and keeping the ear canal open to receive vocal auditory feedback that is coming through the air is one such approach.

8 CONCLUSION

In this paper, we proposed the concept of Digital Speech Makeup (DSM) and empirically showed that speech transformation using voice conversion (VC) can affect one's self-representation in terms of their implicit bias. Our findings and discussions lay the groundwork for the HCI community to work towards designing everyday interpersonal interactions based on the emerging VC techniques, especially in regard to its influence on users' self-representation.

9 ACKNOWLEDGEMENT

This research and development work was supported by JSPS KAKENHI Grant Number JP21J20618 and MIC/SCOPE #182103104.

REFERENCES

[1] Alexander Travis Adams, Jean Marcel dos Reis Costa, Malte F. Jung, and Tanzeem Choudhury. 2015. Mindless computing: designing technologies to subtly influence behavior. In *UbiComp '15*. ACM, New York, NY, 719–730. https://doi.org/10.1145/2750858.2805843

[2] Riku Arakawa, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2019. Implementation of DNN-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device. In *ISCA SSW '19*. ISCA, Singapore, 93–98. https://doi.org/10.21437/ssw.2019-17

[3] Riku Arakawa, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2019. TransVoice: Real-Time Voice Conversion for Augmenting Near-Field Speech Communication. In *UIST '19*. ACM Press, New York, NY, 33–35. https://doi.org/10.1145/3332167.3357106

[4] Riku Arakawa and Hiromu Yakura. 2021. Mindless Attractor: A False-Positive Resistant Intervention for Drawing Attention Using Auditory Perturbation. In *CHI '21*. ACM, New York, NY, 99:1–99:15. https://doi.org/10.1145/3411764.3445339

[5] Jean-Julien Aucouturier, Petter Johansson, Lars Hall, Rodrigo Segnini, Lolita Mercadié, and Katsumi Watanabe. 2016. Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *PNAS* 113, 4 (Jan. 2016), 948–953. https://doi.org/10.1073/pnas.1506552113

[6] D. Banakou, R. Groten, and M. Slater. 2013. Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. *PNAS* 110, 31 (July 2013), 12846–12851. https://doi.org/10.1073/pnas.1306779110

[7] Domna Banakou, Parasuram D. Hanumanthu, and Mel Slater. 2016. Virtual Embodiment of White People in a Black Virtual Body Leads to a Sustained Reduction in Their Implicit Racial Bias. *Front. Hum. Neurosci.* 10 (Nov. 2016). https://doi.org/10.3389/fnhum.2016.00601

[8] Domna Banakou, Sameer Kishore, and Mel Slater. 2018. Virtually Being Einstein Results in an Improvement in Cognitive Task Performance and a Decrease in Age Bias. *Front. Psychol.* 9 (June 2018). https://doi.org/10.3389/fpsyg.2018.00917

[9] Rachel L. Bedder, Daniel Bush, Domna Banakou, Tabitha Peck, Mel Slater, and Neil Burgess. 2019. A mechanistic account of bodily resonance and implicit bias. *Cognition* 184 (March 2019), 1–10. https://doi.org/10.1016/j.cognition.2018.11.010

[10] John W. Black. 1951. The Effect Of Delayed Side-Tone Upon Vocal Rate And Intensity. *J. Speech Hearing Disord.* 16, 1 (March 1951), 56–60. https://doi.org/10.1044/jshd.1601.56

[11] Matthew Botvinick and Jonathan Cohen. 1998. Rubber hands 'feel' touch that eyes see. *Nature* 391, 6669 (Feb. 1998), 756–756. https://doi.org/10.1038/35784

[12] Theresa A. Burnett, Jill E. Senner, and Charles R. Larson. 1997. Voice F0 responses to pitch-shifted auditory feedback: a preliminary study. *J. Voice.* 11, 2 (June 1997), 202–211. https://doi.org/10.1016/s0892-1997(97)80079-3

[13] Maria Christofi and Despina Michael-Grigoriou. 2017. Virtual reality for inducing empathy and reducing prejudice towards stigmatized groups: A survey. In *VSMM'17*. IEEE, New York, NY, 1–8. https://doi.org/10.1109/vsmm.2017.8346252

[14] David M. Corey and Vishnu Anand Cuddapah. 2008. Delayed auditory feedback effects during reading and conversation tasks: Gender differences in fluent adults. *J. Fluency. Disord.* 33, 4 (Dec. 2008), 291–305. https://doi.org/10.1016/j.jfludis.2008.12.001

[15] Richard Corson. 1972. *Fashions in makeup, from ancient to modern times*. Peter Owen Limited.

[16] Jean Costa, Malte F. Jung, Mary Czerwinski, François Guimbretière, Trinh Le, and Tanzeem Choudhury. 2018. Regulating Feelings During Interpersonal Conflicts by Changing Voice Self-perception. In *CHI '18*. ACM Press, New York, NY, 631. https://doi.org/10.1145/3173574.3174205

[17] Donna Z. Davis and Karikarn Chansiri. 2018. Digital identities – overcoming visual bias through virtual embodiment. *Inf. Commun. Soc.* 22, 4 (Nov. 2018), 491–505. https://doi.org/10.1080/1369118x.2018.1548631

[18] Mar Gonzalez-Franco and Tabitha C. Peck. 2018. Avatar Embodiment. Towards a Standardized Questionnaire. *Front. Robot. AI.* 5 (June 2018), 74. https://doi.org/10.3389/frobt.2018.00074

[19] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* 74, 6 (1998), 1464.

[20] Anthony G. Greenwald, Brian A. Nosek, and Mahzarin R. Banaji. 2003. Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *J. Pers. Soc. Psychol.* 85, 2 (2003), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

[21] Victoria Groom, Jeremy N. Bailenson, and Clifford Nass. 2009. The influence of racial embodiment on racial bias in immersive virtual environments. *Soc. Influ.* 4, 3 (July 2009), 231–248. https://doi.org/10.1080/15534510802643750

[22] Fernanda Herrera, Jeremy Bailenson, Erika Weisz, Elise Ogle, and Jamil Zaki. 2018. Building long-term empathy: A large-scale comparison of traditional and virtual reality perspective-taking. *PLOS ONE* 13, 10 (Oct. 2018), e0204494. https://doi.org/10.1371/journal.pone.0204494

[23] Katunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyohiro Shikano, and Shuichi Itahashi. 1999. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *Journal of the Acoustical Society of Japan (E)* 20, 3 (1999), 199–206. https://doi.org/10.1250/ast.20.199

[24] Roz Ivanič and David Camps. 2001. I am how I sound. *J. Second Lang. Writ.* 10, 1-2 (Feb. 2001), 3–33. https://doi.org/10.1016/s1060-3743(01)00034-0

[25] Hideki Kawahara. 1993. Transformed auditory feedback: Effects of fundamental frequency perturbation. *JASA* 94, 3 (Sept. 1993), 1883–1884. https://doi.org/10.1121/1.407536

[26] Konstantina Kilteni, Ilias Bergstrom, and Mel Slater. 2013. Drumming in Immersive Virtual Reality: The Body Shapes the Way We Play. *IEEE Trans. Vis. Comput. Graph.* 19, 4 (April 2013), 597–605. https://doi.org/10.1109/tvcg.2013.29

[27] Rébecca Kleinberger, George Stefanakis, and Sebastian Franjou. 2019. Speech Companions: Evaluating the Effects of Musically Modulated Auditory Feedback on the Voice. In *(ICAD'19)*. Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne. https://doi.org/10.21785/icad2019.035

[28] Bernard S. Lee. 1951. Artificial Stutter. *J. Speech Hearing Disord.* 16, 1 (March 1951), 53–55. https://doi.org/10.1044/jshd.1601.53

[29] Jong-Eun Roselyn Lee, Clifford I. Nass, and Jeremy N. Bailenson. 2014. Does the Mask Govern the Mind?: Effects of Arbitrary Gender Representation on Quantitative Task Performance in Avatar-Represented Virtual Groups. *Cyberpsychol Behav Soc Netw.* 17, 4 (April 2014), 248–254. https://doi.org/10.1089/cyber.2013.0358

[30] Michelle Lincoln, Ann Packman, and Mark Onslow. 2006. Altered auditory feedback and the treatment of stuttering: A review. *J. Fluency. Disord.* 31, 2 (Jan. 2006), 71–89. https://doi.org/10.1016/j.jfludis.2006.04.001

[31] Matthew R. Longo, Friederike Schüür, Marjolein P.M. Kammers, Manos Tsakiris, and Patrick Haggard. 2008. What is embodiment? A psychometric approach. *Cognition* 107, 3 (June 2008), 978–998. https://doi.org/10.1016/j.cognition.2007.12.004

[32] Sarah Lopez, Yi Yang, Kevin Beltran, Soo Jung Kim, Jennifer Cruz Hernandez, Chelsy Simran, Bingkun Yang, and Beste F. Yuksel. 2019. Investigating Implicit Gender Bias and Embodiment of White Males in Virtual Reality with Full Body Visuomotor Synchrony. In *CHI '19*. ACM Press, New York, NY, 557. https://doi.org/10.1145/3290605.3300787

[33] Lara Maister, Natalie Sebanz, Günther Knoblich, and Manos Tsakiris. 2013. Experiencing ownership over a dark-skinned body reduces implicit racial bias. *Cognition* 128, 2 (Aug. 2013), 170–178. https://doi.org/10.1016/j.cognition.2013.04.002

[34] Lynn Carol Miller and Cathryn Leigh Cox. 1982. For Appearances' Sake. *Pers Soc Psychol Bull* 8, 4 (Dec. 1982), 748–751. https://doi.org/10.1177/0146167282084023

[35] Daiki Miyashiro, Akira Toyomura, Tomosumi Haitani, and Hiroaki Kumano. 2019. Altered auditory feedback perception following an 8-week mindfulness meditation practice. *Int. J. Psychophysiol.* 138 (April 2019), 38–46. https://doi.org/10.1016/j.ijpsycho.2019.01.010

[36] Kristine L Nowak and Jesse Fox. 2018. Avatars and computer-mediated communication: A review of the definitions, uses, and effects of digital representations on communication. *Rev. Commun. Res.* 6, 1 (Jan. 2018), 30–53. https://doi.org/10.12840/issn.2255-4165.2018.06.01.015

[37] Veronica S Pantelidis. 2010. Reasons to use virtual reality in education and training courses and a model to determine when to use virtual reality. *Themes in Science and Technology Education* 2, 1-2 (2010), 59–70.

[38] Tabitha C. Peck, Sofia Seinfeld, Salvatore M. Aglioti, and Mel Slater. 2013. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Conscious. Cogn.* 22, 3 (Sept. 2013), 779–787. https://doi.org/10.1016/j.concog.2013.04.016

[39] Laura Rachman, Marco Liuni, Pablo Arias, Andreas Lind, Petter Johansson, Lars Hall, Daniel Richardson, Katsumi Watanabe, Stéphanie Dubal, and Jean-Julien Aucouturier. 2017. DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. *Behav. Res. Methods* 50, 1 (April 2017), 323–343. https://doi.org/10.3758/s13428-017-0873-y

[40] Lisa. E. Rombout and Marie Postma-Nilsenova. 2019. Exploring a Voice Illusion. In *ACII'19*. IEEE, New York, NY, 711–717. https://doi.org/10.1109/acii.2019.8925492

[41] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi. 2018. Non-Parallel Voice Conversion Using Variational Autoencoders Conditioned by Phonetic Posteriorgrams and D-Vectors. In *ICASSP''18*. IEEE, New York, NY, 5274–5278. https://doi.org/10.1109/icassp.2018.8461384

[42] Maria V. Sanchez-Vives, Bernhard Spanlang, Antonio Frisoli, Massimo Bergamasco, and Mel Slater. 2010. Virtual Hand Illusion Induced by Visuomotor Correlations. *PLoS ONE* 5, 4 (April 2010), e10381. https://doi.org/10.1371/journal.pone.0010381

[43] Ulrike Schultze. 2010. Embodiment and presence in virtual worlds: a review. *J. Inf. Technol.* 25, 4 (Dec. 2010), 434–449. https://doi.org/10.1057/jit.2010.25

[44] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2021. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE ACM Trans. Audio Speech Lang. Process.* 29 (2021), 132–157. https://doi.org/10.1109/TASLP.2020.3038524

[45] Yannis Stylianou. 2009. Voice Transformation: A survey. In *ICASSP'09*. IEEE, New York, NY, 3585–3588. https://doi.org/10.1109/icassp.2009.4960401

[46] Yannis Stylianou, Oliver Cappe, and Eric Moulines. 1998. Continuous probabilistic transform for voice conversion. *IEEE Trans. Audio Speech Lang. Process* 6, 2 (March 1998), 131–142. https://doi.org/10.1109/89.661472

[47] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng. 2016. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training. In *ICME'16*. IEEE, New York, NY, 1–6. https://doi.org/10.1109/icme.2016.7552917

[48] Ana Tajadura-Jiménez, Domna Banakou, Nadia Bianchi-Berthouze, and Mel Slater. 2017. Embodiment in a Child-Like Talking Virtual Body Influences Object Size Perception, Self-Identification, and Subsequent Real Speaking. *Sci.* 7, 1 (Aug. 2017). https://doi.org/10.1038/s41598-017-09497-3

[49] Tomoki Toda. 2014. Augmented speech production based on real-time statistical voice conversion. In *GlobalSIP'14*. IEEE, New York, NY, 592–596. https://doi.org/10.1109/globalsip.2014.7032186

[50] Tomoki Toda, Alan W. Black, and Keiichi Tokuda. 2007. Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory. *IEEE Trans. on Audio, Speech and Lang. Processing* 15, 8 (Nov. 2007), 2222–2235. https://doi.org/10.1109/tasl.2007.907344

[51] Tomoki Toda, Takashi Muramatsu, and Hideki Banno. 2012. Implementation of computationally efficient real-time voice conversion. In *ISCA Speech'12*. ISCA, Singapore, 94–97.

[52] Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi. 2016. Multidimensional scaling of systems in the Voice Conversion Challenge 2016. In *ISCA SSW'16*. ISCA, Singapore, 38–43. https://doi.org/10.21437/SSW.2016-7

[53] Aubrey J. Yates. 1963. Delayed auditory feedback. *Psychol. Bull.* 60, 3 (1963), 213–232. https://doi.org/10.1037/h0044155

[54] Nick Yee and Jeremy Bailenson. 2006. Walk a mile in digital shoes: The impact of embodied perspective-taking on the reduction of negative stereotyping in immersive virtual environments. *Presence (Camb)* (01 2006).

[55] Nick Yee and Jeremy Bailenson. 2007. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Hum. Commun. Res.* 33, 3 (July 2007), 271–290. https://doi.org/10.1111/j.1468-2958.2007.00299.x

[56] Zane Z. Zheng, Ewen N. MacDonald, Kevin G. Munhall, and Ingrid S. Johnsrude. 2011. Perceiving a Stranger's Voice as Being One's Own: A 'Rubber Voice' Illusion? *PLoS ONE* 6, 4 (April 2011), e18655. https://doi.org/10.1371/journal.pone.0018655